基于 RPYS i/o 的数字图书馆领域历史根源文献探究*

■ 吴闯 谢福秀 王春蕾 刘万国 孙波
东北师范大学图书馆 长春 130024

摘要：［**目的/意义**］探索某一学科或研究领域的历史根源与演化问题，对其建设与发展具有重要意义。［**方法/过程**］基于可视化在线工具 RPYS i/o 进行标准 RPYS（参考文献出版年图谱）与多维 RPYS 两种分析，发现对数字图书馆领域的起源和演化起到重要影响的文献。［**结果/结论**］该工具提供的标准 RPYS 分析能够较准确的发现该领域起源相关的经典文献，并通过多维 RPYS 分析还可以发现在该研究领域起源过程中起到持久贡献的文献。

关键词：RPYS i/o 数字图书馆 标准 RPYS 多维 RPYS 历史根源
分类号：G253.1
DOI：10.13266/j.issn.0252-3116.2018.05.010

1 引言

“数字图书馆”一词兴起于 20 世纪 90 年代，1993 年由美国国家科学基金会（National Science Foundation，NSF）、美国国防部尖端研究项目机构（Defense Advanced Research Projects Agency，DARPA）、国家航空与太空总署（National Aeronautics and Space Administration，NASA）联合发起数字图书馆创始工程（Digital Library lnitiative，DLI）。1994 年，美国国会图书馆宣布耗巨资建立国家数字图书馆，美国的这一行动推动了世界范围内数字图书馆的建设。数字图书馆前期也被称为电子图书馆、虚拟图书馆，但是关于数字图书馆的概念一直没有统一的说法，数字图书馆发展 20 余载，不仅给图书馆带来了技术层面的改变，更是服务观念与方式的巨大变革，随着计算机技术与现代通讯技术的发展，数字图书馆将成为越来越复杂的研究领域，也将面临更多的问题，因此有必要对数字图书馆领域进行综述研究，理清数字图书馆领域起源及演化发展过程，对于数字图书馆研究的进一步发展具有重要意义。

基于数字图书馆及相关领域文献的调研，发现国内外有文献研究了该领域的起源及发展过程，如邓香莲从电子图书馆与数字图书馆的关系入手，阐述了数字图书馆的起源，并分析归纳了其概念内涵[1]；M. Lesk 从技术、经济、法律及社会等角度回顾了数字图书馆的发展历史及面临的问题[2]，X. Iris 将数字图书馆的发展演化过程划分为 4 个阶段：①1990 年以前的早期幻想时期；②1991－2000 年成长期；③2000－2010 年建设发展及开放获取发展期；④2010 年以后大型数字图书馆的建设时期[3]。以上研究对了解数字图书馆的起源与发展演化过程起到了一定的作用，但上述研究在进行数字图书馆领域综述研究时，均采用人工阅读的系统综述方式，效率低，且容易带有主观性，难以应对海量的学术资源。国内外还有利用共词、共引的文献计量学方法，如杨国立[4]、闫伟东[5]、杨九龙[6]、洪凌子[7]、L. Godeaux[8]等人均采用了该方法对数字图书馆领域的研究现状与热点进行了分析，而对于该领域产生之前的文献即起源问题并未具体研究。

基于以上原因，本文利用新兴的文献计量分析工具 RPYS i/o，来探究数字图书馆领域的历史根源及演化过程，从海量文献中发现可能与该领域起源相关的的经典文献，对于明确数字图书馆的概念内涵以及该领域进一步建设与发展具有重要意义，以期为其他研究者使用该工具寻求其他研究领域的历史根源文献提供借鉴。

* 本文系国家社会科学基金重点项目“基于云计算的国家数字学术信息资源安全保障体系构建研究”（项目编号：14ATQ008）和吉林省社会科学基金项目“利用云计算构建吉林省社会科学数字资源服务平台研究”（项目编号：202214112）研究成果之一。
作者简介：吴闯（ORCID：0000－0002－9499－5541），馆员，硕士，E-mail：wuc203@nenu.edu.cn；谢福秀（ORCID：0000－0003－4690－8703），馆员，硕士；王春蕾（ORCID：0000－0003－0591－6512），副研究馆员，博士；刘万国（ORCID：0000－0002－1917－6407），馆长，研究馆员；孙波（ORCID：0000－0001－9378－4032），副研究馆员，硕士。
收稿日期：2017－09－11 修回日期：2017－11－17 本文起止页码：87－96 本文责任编辑：王传清

2 RPYS i/o 的原理及功能特点

1964 年,在美国空军科学研究办公室(AFOSR)的资助下,E. Garfield、I. H. SHER 和 R . J. TORPIE 讨论了利用引文数据来探索科学领域历史根源的模型与方法[9],2003 年,E. Garfield、A. I. PUDOVKIN 和 V. S. ISTOMIN 基于引文分析开发了引文图谱分析软件 HistCite,可以快速描绘一个学科领域的发展历史,定位该领域的重要文献[10],2013 年,在第十四届国际科学计量学和信息计量学大会(14th International Conference on Scientometrics & In-formatrics,ISSI 2013)上,W. Marx 和 L. Bornmann 首次提出了"参考文献出版年图谱"(Reference Publication Year Spectroscopy,RPYS)这一学科领域历史根源探究的新方法[11],国内研究者李信将 RPYS 总结为"以某一个领域的全部文献所引用的全部参考文献的出版年份(RPYs)为横轴,以每个参考文献出版年(RPY)的全部参考文献的总被引频次为纵轴而形成的二维分布图"[12],RPYS 在算法和可视化方面可以作为 HistCite 方法的补充。

当前 RPYS 分析研究者已经开发出两种软件包:①2014 年,由荷兰阿姆斯特丹大学 L. Leydesdorff 开发的 RPYS. exe,并免费提供给广大研究者使用,其获取网址是 http://www. leydesdorff. net/software/rpys/。②由莱比锡电信应用科技大学 A. Thor 开发的 CRExplor. exe,其软件免费获取网址是 http://andreas-thor. github. io/cre/。CRExplor. exe 相比于 RPYS. exe 多了数据"消歧"功能,能够识别被引用的参考文献的一些变体,即整合由于写法不规范而事实上是同一参考文献的一些数据。

RPYS i/o 是 2016 年由美国弗吉尼亚技术应用研究公司(VTARC)J. A. Comins 与荷兰阿姆斯特丹大学 L. Leydesdorff 共同开发的在线工具,可以进行两种 RPYS 分析[13]:①标准 RPYS(Standard RPYS)分析。标准 RPYS 的原理是从参考文献角度出发,认为在研究领域产生前发表的全部参考文献中,总存在着几篇文献的被引频次远高于同年或前后几年发表的其他文献,这些文献很可能就是对学科领域的起源及演化发挥重要作用的经典文献,而这些文献一定位于图谱的峰值点上,因此,通过对参考文献出版年图谱在学科领域产生之前的引用频次的峰值进行分析,来探索该学科领域的历史根源文献。②多维 RPYS(Multi-RPYS)分析。多维 RPYS 分析的原理是将每年的参考文献进行一次标准 RPYS 分析,即算出参考文献的每年总被引频次相对于前一年、前两年、该被引用年、后一年、后两年的总被引频次的中位数的偏差,再利用秩转换的思想,将偏差数值进行排序,偏差越大则秩的值越高,再将秩值转化为可视化的热度值图谱,热度值越大,颜色越深。因此,热度值图谱的颜色越深,代表偏差值越大,进而表明该 RPY 的被引频次相比与前后几年被引越频繁。多维 RPYS 分析图谱可以表示每一年的参考文献逐年的引用热度及动态变化情况,借此可以一定程度上分析出历史上对该学科或研究领域具有长期贡献的参考文献。

RPYS i/o 与分析软件 RPYS. exe 和 CRExplor. exe 的区别在于:①基于网络在线平台,操作简单,交互性好;②可以进行标准 RPYS 与多维 RPYS 两种分析;③基于 DOI 和 Google 搜索引擎,提供了获取所识别的学科或研究领域的经典文献的链接。而在线工具 RPYS i/o 最大的局限是当前版本仅能够分析大小 15M 以内的数据集,且当前版本可分析的出版年限范围是 1900 – 1999 年。但相比于已有分析软件,该在线工具能够进行多维 RPYS 分析是其最大的优势,此外能够提供获取经典文献的 DOI 也是该工具的特色。

关于 RPYS 的应用效果,国外已有 10 余篇利用 RPYS. exe 或 CRExplor. exe 软件来探究某领域的历史根源的研究,涉及希格斯波色子[14]、石墨烯与太阳能电池[15]、生物学中的"达尔文雀传奇"[16]、全球定位系统[17]及气候变化[18]等各个领域,国内李信、陆伟、李旭晖 2016 年首次利用 RPYS 对健康信息素养领域的历史起源问题进行了研究[19],2017 年,以引文分析[12]和情感分析为例[20],进一步探索了 RPYS 分析作用。上述研究表明 RPYS 在一定程度上揭示一个学科或研究领域的影响深远的重要文献,发现甚至成为无人问津的"睡美人文献"。关于利用 RPYS i/o 在线工具来探究历史根源的文献,国外只有少数几篇,即开发者曾对生物医学的基底细胞癌领域[21]和期刊《journal Philosophy of Science》[13]的根源和经典文献进行了探究,大连大学侯建华老师也利用该工具对引文分析领域的起源问题进行了研究[22],国内中文文献还未见报道。

3 数字图书馆领域历史根源文献探究

3.1 数据来源

笔者选择 Web Of Science(WOS)核心合集作为数据来源,具体包括以下索引:SCI-EXPANDED,SSCI,A&HCI,CPCI-S 及 CPCI-SSH;检索策略为:主题 = "electronic libra * " or "digital libra * " or "virtual libra

＊”;时间跨度为1985年至今,检索时间为2017年8月31日;文献类型选择article,精炼后得到与数字图书馆相关的论文3 621篇,选择“全记录和引用的参考文献”导出格式,一次最多导出500条,再将导出的数个txt文件合并成一个数据文件,将其重命名为data. txt,作为本研究最终分析的数据文件。该数据集大小为13M。

3.2 导入数据

RPYS i/o平台网址是http://comins. leydesdorff. net/,界面见图1,推荐采用Google Crome或Safari浏览器,某些浏览器(例如Firefox)则不太适合运行该工具,将数据文件data. txt上传至该平台即可进行在线分析,需要注意的是该分析工具要求的数据集大小是15M以内。

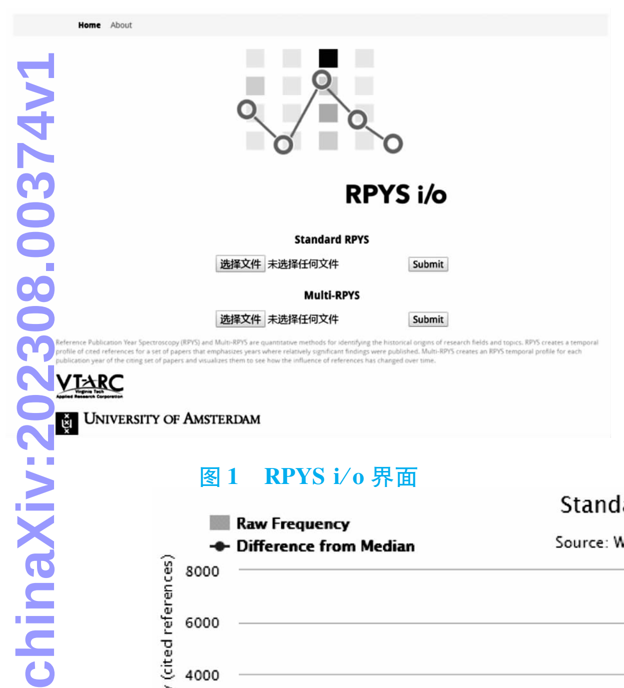


图1　RPYS i/o界面

3.3 标准RPYS运行结果解读

运行结果上方展示的是1900－1999年数字图书馆领域标准参考文献出版年图谱(见图2),该图谱x轴表示的是参考文献出版年,y轴呈现了两组数据值,柱状图表示的是每一年参考文献的被引总频次,样条光滑曲线表示的是该出版年的参考文献的总被引频次相对于该出版年前一年、前两年、该出版年、后一年、后两年的总被引频次的中位数的偏差。鼠标滑过每一出版年,网页上会自动显示该年参考文献的总被引频次及与偏差的具体数值。例如1981－1985年参考文献的总被引频次依次为295、301、470、380、433,可见这5年中位数为380,1983年总被引频次470与该中位数的偏差为90,1983年在样条光滑曲线上呈现出峰值的数值即90。在图谱区域点击并拖动鼠标,可以呈现所选择的年限范围的参考文献出版年图谱,更清晰地展示特定年限范围内曲线及柱形图的变化情况,例如1900－1944年曲线波动不明显,若以1900年为起点拖动鼠标至1944年,平台则重新更加清晰的呈现1900－1945年限范围内的图谱(见图3),若以1946年为起点拖动鼠标至1960年,则也更清晰的呈现该时间段内的峰值情况(见图4)。

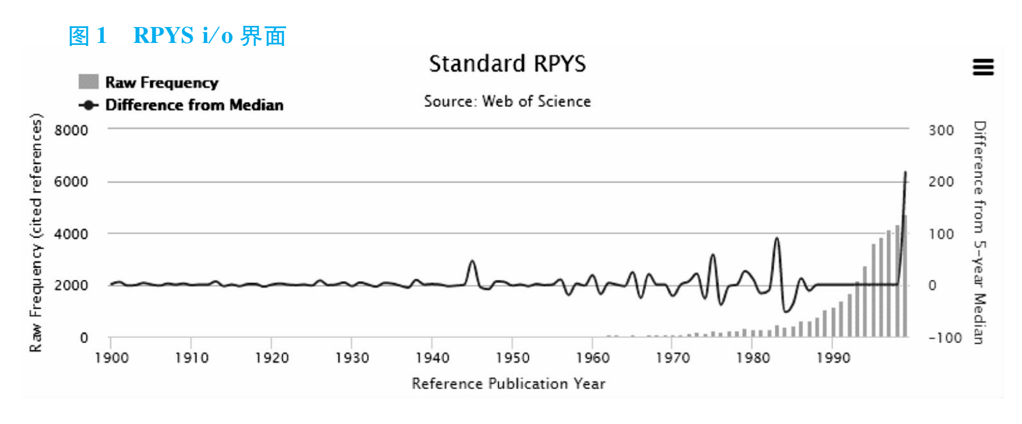


图2　1900－1999年数字图书馆领域标准参考文献出版年图谱

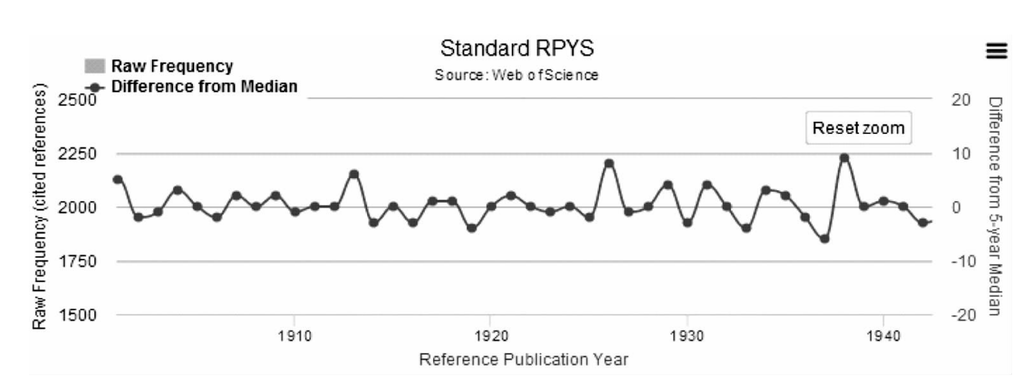


图3　1900－1944年数字图书馆领域标准参考文献出版年图谱

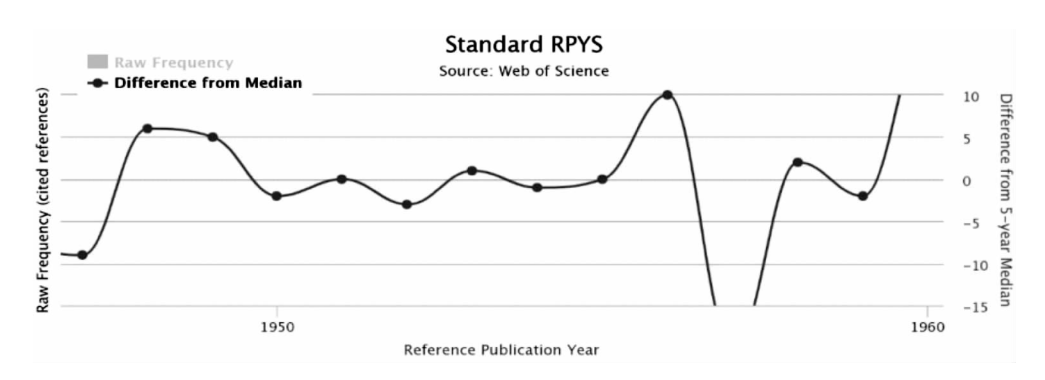


图 4 1946 – 1960 年数字图书馆领域标准参考文献出版年图谱

返回的运行结果页面下方呈现的是可供检索的文献列表(见图 5),数据内容是所检索年份的参考文献按被引频次排名前 40 的文献,该列表包括 5 列内容:第一列为文献作者;第二列为文献的出版年;第三列为该文献的来源出版物;第四列为该文献的总被引频次;第五列为获取该文献的链接。该文献列表可以帮助使用者寻找到该研究领域的重要历史根源文献,如 1983 年呈现出一个明显的峰值,说明该年的参考文献被引用频繁,该年发表的研究成果对数字图书馆领域的研究影响重大,引用频次最高的文献对该领域的起源或发展起到重要作用,在检索框中输入 rpy1983,数据列表给出了出版年 1983 年引用频次排名前 40 的文献,点击 Time Referenced,可以按照被引频次从高到低或从低到高排序。

Search and Refine Results: rpy1983

Author(s)	RPY	Source	Times Referenced	Link
SALTON G	RPY1983	INTRO MODERN INFORMA	37	Try Google Scholar
SALTON G.	RPY1983	INTRO MODERN INFORM	13	Try Google Scholar
CARD S.	RPY1983	PSYCHOL HUMAN COMPUT	8	Try Google Scholar
TUFTE E.	RPY1983	VISUAL DISPLAY QUANT	8	Try Google Scholar
DIMAGGIO PJ	RPY1983	AM SOCIOL REV V48 P147	7	Article Link
SALTON G	RPY1983	COMMUN ACM V26 P1022	7	Article Link
ZADEH LA	RPY1983	COMPUT MATH APPL V9 P149	5	Article Link
JORGENSEN WL	RPY1983	J CHEM PHYS V79 P926	5	Article Link
CULNAN M. J.	RPY1983	DECISION SCI V14 P194	4	Article Link
GOLDBERG A.	RPY1983	SMALLTALK 80 LANGUAG	4	Try Google Scholar
KABSCH W	RPY1983	BIOPOLYMERS V22 P2577	4	Article Link
ANDERSON JR	RPY1983	J VERB LEARN VERB BE V22 P261	4	Article Link

图 5 所检索年份的参考文献按被引频次排名前 40 的文献

3.4 标准 RPYS 结果分析

基于上述图谱结果,对数字图书馆领域的起源与演化起重要作用的文献进行分析,从图 2 可以看出,1900 – 1960 年该领域参考文献年被引总频次整体较低(低于 100 次),1961 – 1985 年该领域的参考文献被引总频次稳步提升(从 100 次左右上升至 400 余次),从 1986 年起该领域参考文献被引总频次一度呈指数型增长(600 余次增加至 4 000 余次),表明数字图书馆领域进入了快速发展时期。基于上述观察结果,本研究将数字图书馆领域产生前 RPY 划分为 1900 – 1960 年、1961 – 1985 年和 1986 – 1993 年 3 个时间段,依据 W. Marx 等 2014 年的研究成果[15],在对 RYYS 峰值点分析时,往往只需要对被引频次最高的单篇文献进行分析,即通过被引频次最高的文献来探究对该学科或领域的起源的重要作用,根据上述 RPYS 分析的原理,观察总被引频次与 5 年总被引频次中位数偏差曲线的峰值点年份,结合文献列表检索出该 RPY 被引频次最高的文献,并进行分析。

3.4.1 1900 – 1960 年数字图书馆领域标准 RPYS 分析 从图 2 可知,1900 – 1960 年间,数字图书馆领域 RPYS 上出现了一个最大的峰值点:1945 年,结合图 3 和图 4,该时间段内也存在一些相对较为明显的峰值点:1913 年、1926 年、1938 年、1956 年和 1960 年,利用文献列表检索出上述 6 个峰值点被引频次最高的文献,见表 1。

由表 1 可见,最大的峰值点处(1945 年)被引频次最多的参考文献是美国科学家 V. Bush(万尼瓦尔. 布什)在《The Atlantic Monthly》(《大西洋月刊》)上发表的一篇文章“As we may think”[23],东北师范大学传媒科学学院徐跃权教授将其翻译为“我们可以这样设想”[24],该文详细的描绘了计算机技术对于科研者在信息收集、存储、发现和检索方面的应用前景。文中 6 次提到图书馆,对图书馆机械化充满了憧憬,布什提出了一个全新的概念“Memex(记忆扩展机)”,将所有资料存于此,一个屏幕、一个键盘、一系列按钮和手柄即可自由阅读。

1913 年峰值点处是提出“记忆遗忘曲线”的德国著名心理学家 H. Ebbinghaus 的论文[25],该文阐述了练习对记忆力的影响研究;1926 年峰值点处是 A. J. Lotka 提出的“洛特卡定律”一文[26],洛特卡定律是文献计量

表 1　1900 –1960 年峰值点被引频次最多的参考文献

RPY	RPY 被引总频次	被引频次最高的参考文献/被引频次
1913	10	EBBINGHAUS H. Memory:a contribution to experimental psychology[M]. Boston:University ,1913./2
1926	11	LOTKA ALFRED J. The frequency distribution of scientific productivity[J]. Journal of the washington academy of sciences,1926,16(12):317 –323./5
1938	16	WELLS H G. World brain[M]. First UK edition. London:Methuen & Co. , 1938./3
1945	321	BUSH V. As we may think[J]. The atlantic monthly,1945,176(1):101 –108. /26
1956	40	MILLER G A. The magical number seven[J]. Psychological review,1956,63(2):81 –97./10
1960	66	COHEN J. A coefficient of agreement for nominal scales[J]. Educational & psychological measurement,1960,20(1):37 –46./16

学中的重要定律,第一次揭示了作者频率与文献数量之间的关系;1938 年峰值点处是 H. G. Wells 的著作《World brain》[27],该文献提出“世界脑”的概念,当作知识的联合系统,所有人都可以访问;1956 年峰值点处是 G. A. Miller 闻名世界的一篇论文,即“神奇的数字 7 ±2”[28],该文指出人的记忆是短时的,人类信息加工能力存在局限;1960 年峰值点处的论文是 J. Cohen 等提出的统计学指标“kappa 系数”[29],该系数作为评价判断一致性程度的指标,在许多研究中被广为应用。

3.4.2　1961 –1985 年数字图书馆领域标准 RPYS 分析　笔者用同样方法观察图 2,可见 1961 –1985 年间有 6 个较为明显的峰值点,分别为:1965 年、1967 年、1973 年、1975 年、1979 年和 1983 年。结合文献列表检索到上述 RPY 被引最高频次的文献,如表 2 所示:

表 2　1961 –1985 年峰值点被引频次最多的参考文献

年份	RPY 被引总频次	被引频次最高的参考文献/被引频次
1965	99	PRICE D J D. Networks of scientific papers[J]. Science,1965,149(3683):510 –515./13
1967	119	GLASER B G, STRAUSS A L. Discovery of grounded theory: strategies for qualitative research[M]. New York:Aldine De Gruyter,1967./15
1973	169	SMALL H. Cocitation in scientific literature - new measure of relationship between 2 documents[J]. Journal of the American Society for Information Science,1973,24(4):265 –269./18
1975	236	SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J]. Communications of the ACM,1975,18(11):613 –620./21
1979	321	VAN RIJSBERGEN C J. Information retrieval[M]. London:Butterworths,1979./26
1983	470	SALTON G. Introduction to modern information retrieval[M]. New York:McGraw-Hill,1983./37

该时间阶段内第一个峰值点被引频次最高的参考文献是 D. J. D. PRICE(普赖斯)1965 年发表在《Science》期刊中的“Networks of scientific papers”一文[30],普赖斯是美国著名科学学家、科学史家,情报科学的创始人之一,科学计量学之父。这篇论文是普赖斯发表论文中最著名的一篇,他利用 SCI 的引文数据,展示了科学论文之间的引证关系、引证网络以及这种关系和网络可以如何用来进行科学计量学的研究;第二峰值点文献是 B. G. Glaser 和 A. L. Strauss 于 1967 年共同提出了一种定性研究方法“扎根理论”一文[31],扎根理论的首要任务是建立介于宏大理论和微观操作性假设之间的实质理论(即适用于特定时空的理论),但也不排除对具有普适性的形式理论的建构,该理论在图书情报领域也被广为应用,如王平[32]、林婷[33]、茹嘉祎[34]等基于扎根理论对图书馆服务与管理进行了探讨;第三峰值点文献是美国的 H. Small 于 1973 年提出了共引分析的概念一文[35],引文分析中最具影响力的就是共引分析方法,用于揭示科学研究现状、演化及前沿等问题的研究;第四个峰值点文献是 G. Salton 发表于 1975 年的著名成果“IR 向量空间模型”[36],G. Salton 被公认为是现代搜索技术之父,现代信息检索的奠基人,IR 向量空间模型成功地应用于了著名的 SMART 文本检索系统中;第五个峰值点指向的是 C. J. Van Rijsbergen 的专著《Information Retrieval》[37],C. J. Van Rijsbergen 本人被公认为现代信息检索的创始人之一,《Information Retrieval》则被认为是信息检索方面经典的教科书;第六个峰值点文献指向的是 G. Salton 的《Introduction to Modern Information Retrieval》一书[38],为情报检索提供了理论的基础,被广为引用。

3.4.3　1986 –1993 年数字图书馆领域标准 RPYS 分析　从图 2 中可以发现,此时间段的图谱曲线特点:1986 年是个高峰期,1987 年处于相对低谷期,1988 年 –1993 年均为直线,即中位数为 0,表示 1988 年 –1993 年间每个 RPY 的参考文献总被引频次相对于前后 5 年内的中位数没有提升也没有下降,这说明 RPYS 在分析 RPY 的总被引频次在连续增长的情况下是失

效的，但从图 2 中可以发现参考文献总被引频次呈现逐年快速增长趋势，说明数字图书馆领域的文献进入了快速发展时期，为 1994 年数字图书馆领域的兴起起到了推波助澜的作用，每一年被引频次最高的文献仍是值得关注的经典文献，文献详细信息如表 3 所示：

表 3　1986 – 1993 年历年被引频次最多的参考文献

年份	RPY 被引总频次	被引频次最高的参考文献/被引频次
1986	625	BATES M J. Subject access in online catalogs: a design model[J]. Journal of the Association for Information Science & Technology, 1986, 37(6):357 – 376. /13
1987	614	FURNAS G W, LANDAUER T K, GOMEZ L M, et al. The vocabulary problem in human-system communication[J]. Communications of the ACM, 1987, 30(11):964 – 971.
1988	781	SALTON G, BUCKLEY G. Term-weighting approaches in automatic text retrieval[J]. Information processing & management, 1988, 24(5):513 – 523. /29
1989	1 049	DAVIS F D. Perceived usefulness, perceived ease of use, and user acceptance of information technology[J]. Society for Information Management and the Management Information Systems Research Center, 1989, 13(3):319 – 340. /45
1990	1 165	DEERWESTER S, DUMAIS S T, FURNAS G W, et al. Indexing by latent semantic analysis[J]. Journal of the Association for Information Science & Technology, 1990, 41(6):391 – 407. /24
1991	1 400	KUHLTHAU C C. Inside the search process: information seeking from the user's perspective[J]. Journal of the Association for Information Science & Technology, 1991, 42(5):361 – 371. /18
1992	1 690	GOLDBERG D. Using collaborative filtering to weave an information tapestry[J]. Communications of the ACM, 1992, 35(12):61 – 70. /19
1993	2 189	FOX E A, HIX D, NOWELL L T, et al. Users, user interfaces, and objects: Envision, a digital library[J]. Journal of the American Society for Information Science, 1993, 44(8):480 – 491. /22

由表 3 可见，该时间段内单篇最高被引频次论文集中于信息检索方面，表 3 中第一篇是 M. J. Bates 发表于 1986 年的论文[39]，研究了如何构建合适的查询模型，满足信息检索的需求，第二篇是 G. W. Furnas 于 1987 年发表在《Communications of the ACM》杂志上的论文[40]，讨论了信息检索的词汇匹配问题；第三篇是 G. Salton 等于 1988 年发表的"Term-weighting approaches in automatic text retrieval"一文[41]，文章总结了关于自动加权的观点，并提出了自动文本检索的加权方法；第四篇是美国学者 F. D. Davis 关于用户对信息系统接受的模型的研究[42]，其目的是找出一种有效的行为模式，用于解释使用者对新的信息系统接受的行为；第五篇是 S. Deerwester 于 1990 年建立的基于语义的自动文献标引和检索方法一文[43]；第六篇文献的作者是 C. C. Kuhlthau[44] 从认知角度出发，解释信息检索过程就是一种认知过程，属于信息行为学研究范畴；第七篇文献的作者是 D. Goldberg[45]，文中最先提出了"协同过滤"概念，并将其运用到 Tapestry 系统，协同信息推荐系统逐渐被应用到数字图书馆领域中，并成为该领域的主要研究主题之一；第八篇是 E. Fox 的研究[46]，E. Fox 是数字图书馆领域的先驱，该文献对数字图书馆可用性评价进行了研究。

3.5　多维 RPYS 运行结果解读

上传数据集 data. txt 进行 Multi-RPYS 分析，图 6 即为 1900 – 1999 年数字图书馆领域多维参考文献出版年图谱，该图谱 x 轴表示的是参考文献出版年，y 轴的颜色表示每年参考文献被引用的热度值，热度值越大，颜色越深。若呈现明显连续颜色较深的条带，则表明该年参考文献在此期间被持续引用，借此可进一步探究该年被持续引用的重要文献。

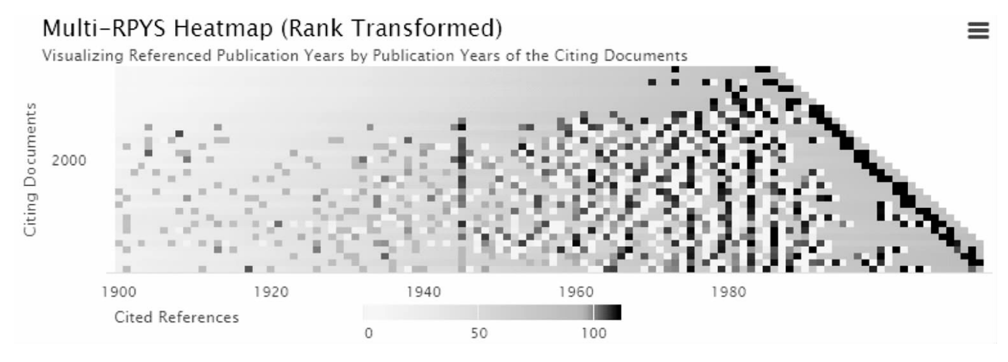


图 6　1900 – 1999 年数字图书馆领域多维参考文献出版年图谱

吴闯，谢福秀，王春蕾，等. 基于 RPYS i/o 的数字图书馆领域历史根源文献探究[J]. 图书情报工作,2018,62(5):87－96.

运行结果页面下方呈现的是可供检索的文献列表(见图 7),可以按出版年检索(也可按作者和期刊进行检索),数据内容是所检索年份的参考文献历年的被引用频次,该列表包括 6 列内容,第一列为文献作者,第二列为文献出版年,第三列为文献来源出版物,第四列为该文献的被引频次,第五列为被引用年份,第六列为获取该文献的链接。

Search and Refine Results: rpy1945

Author(s) ▲	RPY	Source	Times Referenced	CPY	Link
BERNAL JD	RPY1945	P 20 C ASL LOND	1	CPY2000	Try Google Scholar
BOGGS SAMUEL WHITTEMORE	RPY1945	CLASSIFICATION CATAL	1	CPY1998	Try Google Scholar
BUSH V	RPY1945	ATLANTIC V176 P102	1	CPY1994	Try Google Scholar
BUSH V	RPY1945	SCIENCE V102 P79	1	CPY2010	Try Google Scholar
BUSH V	RPY1945	LIFE V19 P121	1	CPY2010	Try Google Scholar
BUSH V	RPY1945	AS WE MAY THINK JUL	1	CPY1999	Try Google Scholar
BUSH V	RPY1945	LIFE V19 P112	1	CPY2010	Try Google Scholar
BUSH V	RPY1945	LIFE V19 P116	1	CPY2010	Try Google Scholar
BUSH V.	RPY1945	ATLANTIC MONTHLY V176 P101	1	CPY2012	Try Google Scholar
BUSH V.	RPY1945	ATLANTIC MONTHLY V176 P101	1	CPY1998	Try Google Scholar
BUSH V.	RPY1945	ATLANTIC MONTHLY V176 P101	2	CPY1995	Try Google Scholar
BUSH V.	RPY1945	ATLANTIC MONTHLY V176 P101	2	CPY2003	Try Google Scholar
BUSH V.	RPY1945	ATLANTIC MONTHLY	1	CPY1998	Try Google Scholar
BUSH V.	RPY1945	ATLANTIC MONTHLY V176 P101	1	CPY2010	Try Google Scholar
BUSH V.	RPY1945	LIFE V19 P123	1	CPY2010	Try Google Scholar
BUSH V.	RPY1945	ATLANTIC MONTHLY JUL	1	CPY1997	Try Google Scholar
BUSH V.	RPY1945	ATLANTIC MONTHLY V176 P101	1	CPY2011	Try Google Scholar

图 7　1945 年参考文献历年被引用频次检索结果

3.6　多维 RPYS 运行结果分析

从图 6 中可以发现,3 个出版年有较为明显的连续深色条带,第一个是 1945 年出版的参考文献在 1994－2006 年间所呈现出明显较深的条带,表明该年参考文献在此期间被持续引用。第二个是 1975 年出版的参考文献在 1999－2004 年、2008－2012 年与 2013－2017 年间所呈现出明显较深的条带,第三个是 1983 年出版的参考文献在 1999－2004 年与 2006－2016 年间所呈现出明显较深的条带。

在数据列表中检索 RPY 为 1945 年的参考文献,可以观察到 V. Bush 的“As we may think”一文,在 1994－2006 年间,逐年被连续引用,产生了持续的影响力(见图 7)。

笔者采用同样方法可以获得 1975 年的参考文献(见图 8),发现 G. Salton 发表于 1975 年的著名成果“IR 向量空间模型”一文,IR 向量空间模型作为信息检索中的最基本的方法之一,在 90 年代数字图书馆兴起以后,历年都被诸多文献引用,当然该 RPY 产生持续影响的论文不仅这一篇,以被引频次排序,还可以发现更多较为重要的文献,本文不再赘述。

采用同样方法获得 1983 年的参考文献(见图 9),发现 G. Salton 的《Introduction to Modern Information Retrieval》(《现代情报检索导论》)一书在 1997 年以后的数年里均被多次引用,可见该文不仅是对数字图书馆的起源起到了举足轻重的地位,对数字图书馆长期发展历史上也存在持久的重要贡献。必须值得指出的是图 6 的右侧斜坡区域,可以看到在斜坡的边缘处颜色均较深,说明文献在出版年的头两年或三年内热度值较高,比较容易受到关注和引用,呈现出短期内被引用较频繁。

3.7　研究结论

基于标准的 PPYS 与多维 RPYS 分析结果,本文共发现了 20 篇对数字图书馆领域起源有重要贡献的文献,还发现其中的 3 篇文献对数字图书馆领域的演化发展产生了持续影响,其中有 8 篇文献集中发表在《Journal of the Association for Information Science & Technology》

SALTON G	RPY1975	COMMUN ACM V18 P613	3	CPY2012	Article Link
SALTON G	RPY1975	COMMUN ACM V18 P613	3	CPY2015	Article Link
SALTON G	RPY1975	COMMUN ACM V18 P613	2	CPY2009	Article Link
SALTON G	RPY1975	COMMUN ACM V18 P613	1	CPY2010	Article Link
SALTON G	RPY1975	COMMUN ACM V18 P613	1	CPY1998	Article Link
SALTON G	RPY1975	COMMUN ACM V18 P613	1	CPY2006	Article Link
SALTON G	RPY1975	COMMUN ACM V18 P613	1	CPY2013	Article Link
SALTON G	RPY1975	COMMUN ACM V18 P613	2	CPY2002	Article Link
SALTON G	RPY1975	COMMUN ACM V18 P613	1	CPY2008	Article Link
SALTON G	RPY1975	COMMUN ACM V18 P613	2	CPY2011	Article Link
SALTON G	RPY1975	COMMUN ACM V18 P613	1	CPY2000	Article Link
SALTON G	RPY1975	COMMUN ACM V18 P613	1	CPY2014	Article Link
SALTON G	RPY1975	COMMUN ACM V18 P613	2	CPY2001	Article Link
SALTON G	RPY1975	THEORY OF INDEXING	2	CPY2016	Try Google Scholar

图 8 1975 年参考文献历年被引用频次检索结果

Search and Refine Results: rpy1983

Author(s)	RPY	Source	Times Referenced	CPY	Link
SALTON G	RPY1983	INTRO MODERN INFORMA	8	CPY2003	Try Google Scholar
SALTON G	RPY1983	INTRO MODERN INFORMA	6	CPY1997	Try Google Scholar
SALTON G	RPY1983	INTRO MODERN INFORMA	4	CPY2000	Try Google Scholar
SALTON G	RPY1983	INTRO MODERN INFORMA	3	CPY1999	Try Google Scholar
SALTON G.	RPY1983	INTRO MODERN INFORM	3	CPY2010	Try Google Scholar
SALTON G.	RPY1983	INTRO MODERN INFORM	3	CPY2008	Try Google Scholar
SALTON G	RPY1983	INTRO MODERN INFORMA	3	CPY2004	Try Google Scholar
TUFTE E.	RPY1983	VISUAL DISPLAY QUANT	3	CPY2001	Try Google Scholar
SALTON G	RPY1983	INTRO MODERN INFORMA	3	CPY2002	Try Google Scholar
SALTON G	RPY1983	INTRO MODERN INFORMA	3	CPY2001	Try Google Scholar
SMEATON AF	RPY1983	COMPUT J V26 P239	2	CPY2002	Article Link
SALTON G	RPY1983	INTRO MODERN INFORMA	2	CPY2005	Try Google Scholar

图 9 1983 年参考文献历年被引用频次检索结果

《Journal of the Association for Information Science & Technology》《Journal of the American Society for Information Science》这 3 种期刊上，可见这 3 种期刊对数字图书馆领域的起源较为重要，其中有 3 篇文献均为同一作者 G. Salton，证实了 G. Salton 对数字图书馆领域起源及发展的重要地位。结合国内外数字图书馆领域的综述文章，笔者对数字图书馆起源的演化过程有了新的认识，并将其定义为 3 个阶段：第一个阶段是幻想时期（1913 – 1960 年），H. Ebbinghaus、G. A. Mille 指出了短时记忆的局限，H. G. Wells 与 V. Bush 对可进行长期大量存储并易于获取的系统或设备进行了想象。第二个阶段是奠基时期（1961 – 1985 年），该时期积累了一些信息检索和科学计量的基础研究，为数字图书馆领域的萌芽奠定了基础。如 D. J. D. PRICE 1965 年一文和 H. Small 1973 年一文均为科学计量史上的著名论文，G. Salton 提出的 IR 向量模型是文本检索系统的基础，C. J. Van Rijsbergen1979 年一书和 G. Salton1983 年一书是信息检索领域的经典教材。第三个阶段是发展时期（1986 – 1994 年），该时间段仍集中于信息检索方向，但研究的问题更加细化和深入，推动了数字图书馆领域快速兴起，例如 M. J. Bates 讨论了如何建立查询模型，G. W. Furnas 讨论了信息检索的词

汇匹配问题,G. Salton 提出了自动文本检索的加权方法,S. Deerwester 建立了基于语义的自动文献标引和检索方法,此阶段的研究还倾向于研究数字图书馆与用户交互问题,该问题后续也被研究者尤为关注,广为研究。综上所述,通过分析对数字图书馆领域产生前的重要起源文献和演化过程有了新的发现,这是很多综述文献都未做到的。

4 结语

本文利用可视化工具 RPYS i/o,探索并发现对数字图书馆领域的起源和演化起到重要影响的文献,该方法能够较准确地发现该领域起源相关的经典文献,并通过可视化的图谱探究在该领域发展过程中起到持久贡献的文献,这些文献在数字图书馆领域的起源与演化中起到了举足轻重的作用,但仍需注意的是,要确定这些文献是否为该学科领域的根源文献,还需要经过该领域专业人员的分析与认定,该工具在分析数据集大小及年代上还存在一定的限制。

参考文献:

[1] 邓香莲. 数字图书馆研究起源及概念内涵分析[J]. 图书馆工作与研究,2003(1):18－20.
[2] LESK M. A personal history of digital libraries[J]. Library hi tech,2012,30(4):592－603.
[3] IRIS X, KRYSTYNA K. Chapter 1-Introduction to digital libraries, in discover digital libraries[M]. Oxford :Elsevier,2016.
[4] 杨国立. 国外数字图书馆研究进展:基于关键词共现和文献共被引的可视化研究[J]. 图书馆杂志,2012(6):20－25.
[5] 闫伟东. 数字图书馆发展的可视化分析[J]. 公共图书馆,2012(1):30－34.
[6] 杨九龙,杜文龙. 基于知识可视化图谱的国内数字图书馆研究演进分析[J]. 图书馆学研究,2012(5):5－9,39.
[7] 洪凌子,黄国彬,于洋. 基于 CiteSpace 的国内外数字图书馆研究论文的比较分析[J]. 图书馆论坛,2014(6):91－100.
[8] GODEAUX L. A co-word analysis of digital library field in China[J]. Scientometrics, 2012, 91(1):203－217.
[9] GARFIELD E, SHER I H, TORPIE R . J. The use of citation data in writing the history of science[M]. Philadelphia:Institute for Scientific Information,1964.
[10] GARFIELD E, PUDOVKIN A I, ISTOMIN V S. Why do we need algorithmic historiography? [J]. Journal of the American Society for Information Science and Technology,2003,54(5):400－412.
[11] MARX W, BORNMANN L, BARTH A. Detecting the historicalroots of research fields by reference publication year spectroscopy (RPYS)[C]//Proceedings of 14th International Society of Scientometrics and Informetrics conference. Stolber - gasse: Facultasv, 2013:493－506.
[12] 李信,李倩. 传统文献计量与科学评价的一个补充视角:全时间域的 RPYS[J]. 图书情报知识,2017(4):89－99.
[13] COMINS J A, LEYDESDORFF L. RPYS i/o: software demonstration of a web-based tool for the historiography and visualization of citation classics, sleeping beauties and research fronts[J]. Scientometrics,2016,107(3):1509－1517.
[14] MARX W, BORNMANN L. On the origins and the historical roots of the Higgs boson research from a bibliometric perspective[J]. European physical journal plus,2014,129(6):1－13.
[15] MARX W, BORNMANN L, BARTH A. Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS)[J]. Journal of the Association for Information Science and Technology,2014,65(4):751－764.
[16] MARX W, BORNMANN L. Tracing the origin of a scientific legend by reference publication year spectroscopy (RPYS) : the legend of the Darwin finches[J]. Scientometrics,2014,99(3):839－844.
[17] COMINS J A, HUSSEY T W. Detecting seminal research contributions to the development and use of the global positioning system by reference publication year spectroscopy[J]. Scientometrics,2015,104(2): 575－580.
[18] MARX W, HAUNSCHILD R, THOR A. Which early works are cited most frequently in climate change research literature? A bibliometric approach based on Reference Publication Year Spectroscopy [J]. Scientometrics,2017,110(1):335－353.
[19] 李信,陆伟,李旭晖. 一种新兴的学科领域历史根源探究方法:RPYS[J]. 图书情报工作,2016,60(20):70－76.
[20] 李信,赵薇,肖香龙,等. 基于 RPYS 分析的引文分析研究:起源和演化[J]. 图书馆论坛,2017, 37(11):56－65.
[21] COMINS J A, LEYDESDORFF L. Citation algorithms for identifying research milestones driving biomedical innovation[J]. Scientometrics,2017,110(3): 1495－1504.
[22] HOU J. Exploration into the evolution and historical roots of citation analysis by referenced publication year spectroscopy[J]. Scientometrics,2017,110(3): 1437－1452.
[23] BUSH V. As we may think[J]. The atlantic monthly,1945,176(1):101－108.
[24] 徐跃权,于宁. 科技名篇《As We May Think》解读[J]. 图书馆杂志,2006,25(11):11－14.
[25] EBBINGHAUS H. Memory:a contribution to experimental psychology[M]. Boston:University Microfilms,1913.
[26] LOTKA ALFRED J. The frequency distribution of scientific productivity[J]. Journal of the Washington Academy of Sciences, 1926,16(12):317－323.
[27] WELLS H G. World brain[M]. First UK edition. LondonMethuen & Co. , 1938.
[28] MILLER G A. The magical number seven[J]. Psychological review,1956,63:81－97.
[29] COHEN J. A coefficient of agreement for nominal scales[J]. Edu-

cational & psychological measurement,1960,20(1):37 -46.

[30] PRICE D J D. Networks of scientific papers[J]. Science,1965,149(3683):510 -515.

[31] GLASER B G, STRAUSS A L. Discovery of grounded theory: strategies for qualitative research [M]. New York: Aldine De Gruyter,1967.

[32] 王平,茹嘉祎. 国内未成年人图书馆服务满意度影响因素——基于扎根理论的探索性研究[J]. 图书情报工作,2015,59(19):41 -46.

[33] 林婷. 基于经典扎根理论的我国高校图书馆 Folksonomy 管理机制实证研究[J]. 图书情报工作,2015,59(16):60 -67.

[34] 柯平,张文亮,李西宁,等. 基于扎根理论的馆员对公共图书馆组织文化感知研究[J]. 中国图书馆学报,2014,40(3):37 -49.

[35] SMALL H. Cocitation in scientific literature-new measure of relationship between 2 documents [J]. Journal of the American Society for Information Science,1973,24(4):265 -269.

[36] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J]. Communications of the ACM,1975,18(11):613 -620.

[37] VAN RIJSBERGEN C J. Information retrieval[M]. London:Butterworths,1979.

[38] SALTON G. Introduction to modern information retrieval[M]. New York:McGraw-Hill,1983.

[39] BATES M J. Subject access in online catalogs: a design model [J]. Journal of the Association for Information Science & Technology,1986,37(6):357 -376.

[40] FURNAS G W, LANDAUER T K, GOMEZ L M, et al. The vocabulary problem in human-system communication[J]. Communications of the ACM, 1987, 30(11):964 -971.

[41] SALTON G, BUCKLEY G. Term-weighting approaches in automatic text retrieval[J]. Information processing & management,1988,24(5):513 -523.

[42] DAVIS F D. Perceived usefulness, perceived ease of use, and user acceptance of information technology[J]. Society for Information Management and the Management Information Systems Research Center,1989,13(3):319 -340.

[43] DEERWESTER S, DUMAIS S T, FURNAS G W, et al. Indexing by latent semantic analysis[J]. Journal of the Association for Information Science & Technology,1990,41(6):391 -407.

[44] KUHLTHAU C C. Inside the search process: information seeking from the user's perspective[J]. Journal of the Association for Information Science & Technology,1991,42(5):361 -371.

[45] GOLDBERG D. Using collaborative filtering to weave an information tapestry[J]. Communications of the ACM,1992,35(12):61 -70.

[46] FOX E A, HIX D, NOWELL L T, et al. Users, user interfaces, and objects: envision, a digital library[J]. Journal of the American Society for Information Science,1993,44(8):480 -491.

作者贡献说明:

吴闯:设计论文研究思路与框架,撰写论文;
谢福秀:参与研究思路设计,修改论文部分内容;
王春蕾:参与研究思路设计,修改论文部分内容;
刘万国:给予论文整体思路指导与修改意见;
孙波:修改论文部分内容。

Detecting the Historical Root of Digital library Research Area Based on RPYS i/o

Wu Chuang Xie Fuxiu Wang Chunlei Liu Wanguo Sun Bo
Northeast Normal University Library, Changchun 130024

Abstract: [**Purpose/significance**]This article aims to explore seminal works about the historical roots of a specific research field or subject. The study of historical roots is of great significance for the construction and research. [**Method/process**]We describe a technical advancement for developing research historiographies by introducing RPYS i/o, an online tool for performing standard RPYS and multi-RPYS analyses. Based on RPYS i/o, we take digital library research field as an example. [**Result/conclusion**]The tool enables users to explore seminal works underlying a research field and to plot the influence of these seminal works over time.

Keywords: RPYS i/o digital library standard RPYS multi-RPYS historical roots